2024 届本科生学士学位论文　　学校代码：10269

華東師範大學

East China Normal University

本科生毕业论文

金融舆情事件检测技术及系统实现

Financial Sentiment Event Detection Techniques and System Implementation

姓　　名：汤应达

学　　号：10205102476

学　　院：计算机科学与技术

专　　业：计算机科学与技术

指导教师：王晓玲

职　　称：教授

2024 年 5 月

华东师范大学学位论文诚信承诺

本毕业论文是本人在导师指导下独立完成的，内容真实、可靠。本人在撰写毕业论文过程中不存在请人代写、抄袭或者剽窃他人作品、伪造或者篡改数据以及其他学位论文作假行为。

本人清楚知道学位论文作假行为将会导致行为人受到不授予/撤销学位、开除学籍等处理（处分）决定。本人如果被查证在撰写本毕业论文过程中存在学位论文作假行为，愿意接受学校依法作出的处理（处分）决定。

承诺人签名：　　日期：　2024 年　　月　　日

华东师范大学学位论文使用授权说明

本论文的研究成果归华东师范大学所有，本论文的研究内容不得以其它单位的名义发表。本学位论文作者和指导教师完全了解华东师范大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权华东师范大学可以将论文的全部或部分内容编入有关数据库进行检索、交流，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

保密的毕业论文（设计）在解密后应遵守此规定。

作者签名：　　导师签名：　　日期：　2024 年　　月　　日

目录

金融舆情事件检测技术及系统实现

摘要：

金融市场是一个受多种因素影响的复杂系统，包括宏观经济数据、政治事件和市场情绪等。随着互联网的发展，大量金融信息和用户观点在网络上迅速传播，这些信息中的舆情事件对金融市场有重大影响。特别是突发金融事件，会引起公众广泛关注，进而对金融市场产生现实影响。因此，金融机构、投资者和市场监管机构需实时监控市场舆情，以快速响应市场变化，控制投资机会或防范风险。同时，有效的金融舆情事件检测也有利于维护网络和社会秩序，保障市场和社会稳定。

本论文的主要研究目的是在之前的研究基础上，基于预训练的语言模型（PLM）完善并改进一个金融舆情事件检测系统。该系统旨在监测和分析网络和社交媒体上的金融信息，以便快速识别可能对金融市场产生重大影响的舆情事件，如企业突发新闻、市场谣言、监管政策变化等。本研究首先通过从诸如 Twitter 这样的社交媒体平台上获取与指定关键词相关的推文数据，并且对这些数据进行预处理和简要的分析。之后，在系统中使用 BART-Large-CNN 模型完成突发事件的检测和事件摘要、总结的生成，并且通过使用 PEGASUS-newsroom 和 T5 这两个模型完成的对比实验说明了 BART-Large-CNN 在本研究的任务上表现更优越，因此最终选择该模型作为系统的核心。

综上所述，本研究完善的系统主要利用预训练语言模型强大的自然语言理解能力、文本分析能力和摘要生成能力，精准地从海量文本数据中抽取和归类相关事件，从而为金融机构、投资者和市场监管机构提供实时的市场情报和风险预警。

关键词：金融舆情事件检测，预训练语言模型，BART-Large-CNN

Financial Sentiment Event Detection Techniques and System Implementation

Abstract:

The financial market is a complex system influenced by various factors, including macroeconomic data, political events, and market sentiment. With the development of the internet, a large amount of financial information and user opinions spread rapidly online, and these public sentiment events have a significant impact on financial markets. Particularly, sudden financial events attract widespread public attention and thus have real effects on financial markets. Therefore, financial institutions, investors, and market regulators need to monitor market sentiments in real time to quickly respond to market changes, seize investment opportunities, or guard against risks. Effective detection of financial sentiment events is also crucial for maintaining network and social order, and ensuring market and social stability.

The main research objective of this thesis is to improve and refine a financial sentiment event detection system based on pre-trained language models (PLMs), building on previous research. This system aims to monitor and analyze financial information on the internet and social media, in order to rapidly identify sentiment events that could significantly impact financial markets, such as corporate breaking news, market rumors, and changes in regulatory policies. This study first collects tweet data related to specified keywords from social media platforms like Twitter, and preprocesses and conducts preliminary analyses of these data. Subsequently, the system uses the BART-Large-CNN model to detect sudden events and generate event summaries and analyses, and comparative experiments conducted with the PEGASUS-newsroom and T5 models demonstrate the superior performance of BART-Large-CNN in the tasks of this study, thus it was chosen as the core of the system.

In summary, the refined system in this study primarily utilizes the powerful natural language understanding, text analysis, and summary generation capabilities of pre-trained language models to accurately extract and categorize relevant events from massive text data, thereby providing real-time market intelligence and risk alerts to financial institutions, investors, and market regulators.

Keywords: Financial Sentiment Event Detection, Pre- trained Language Models, BART-Large-CNN

1、绪论

1.1 研究背景与意义

1.1.1 研究背景

金融市场的动态性和复杂性意味着它是由无数交织的因素和事件驱动的，其中，宏观经济指标、企业财务报告、政治变动以及市场参与者的情绪等都扮演着至关重要的角色。这些因素的变化能够迅速通过互联网和社交媒体平台传播开来，进而影响投资者的决策和市场的整体走向。在这个信息爆炸的时代，区分和识别哪些信息是具有市场影响力的舆情事件变得尤为关键。特别是那些能够引起广泛公众关注和讨论的突发金融事件，如全球性的健康危机（例如新冠疫情）、重大的经济政策调整（如注册制改革）等，它们不仅能够在短时间内极大地改变市场情绪，还可能对金融市场的短期波动和长期走向产生深远影响。因此，实时监测和分析这些舆情事件对于理解市场动态、指导投资决策和维护市场稳定至关重要。

互联网和社交媒体的普及大大增加了信息的可获取性，但同时也带来了信息过载和信息真实性的挑战。在这些平台上，不仅有大量的实时金融信息和数据，还有来自各方的观点和评论，包括专业分析师、业余投资者、甚至是有意图的操纵者。在这种环境下，一些商业网站和自媒体账号可能会发布未经验证的财经新闻、歪曲的经济政策解读、或是对金融市场进行无根据的炒作和贬低，这些行为不仅误导公众，还可能引起市场恐慌或不必要的乐观，对金融市场的稳定构成威胁。

1.1.2 研究意义

在上述背景下，金融领域的事件检测逐渐成为研究者和投资人员共同关注的焦点。高效、准确的金融舆情事件检测技术不仅可以减少人工筛选工作量，节省劳动力，还能帮助提高信息处理的效率。对于金融机构和投资者而言，这种技术有助于进行风险管理和投资决策，尤其是在应对市场快速变化时，实时监控市场舆情成为获取投资优势和防范风险的关键。

随着大数据和人工智能技术的进步，特别是语言模型在金融领域的应用，金融舆情事件检测的研究和实践已取得显著进展。通过利用先进的自然语言处理（Natural Language Processing）技术和机器学习算法，研究人员能够开发出更为精准和高效的舆情事件检测系统。这些系统不仅能够识别和分析金融新闻、社交媒体等来源的文本数据，还能够预测其对金融市场的潜在影响，为市场参与者提供有价值的洞察。

因此，研究和开发金融舆情事件检测技术及系统不仅是维护网络传播秩序的需要，更是保护投资者利益、维护金融市场稳定和促进健康经济发展的重要措施。通过有效的监测和分析系统，可以及时识别和过滤掉误导性和有害的信息，同时将关键的市场影响事件快速传递给市场参与者，帮助他们做出更加明智和信息充分的决策。此外，这样的系统还能为政府监管机构提供实时数据支持，使其能够在必要时采取措施，减轻不利事件对市场的影响，维护金融市场的正常运行和公众信心。这不仅对金融市场的健康运行至关重要，也对整个社会的稳定与发展具有深远的影响。通过提高金融舆情检测的准确性和效率，可以增强市场透明度和稳定性，进而促进经济的持续健康发展。

1.1.3 研究基础

本研究在之前研究的基础上，对 EDAS[1]系统进行了改进和完善。首先是重新搭建了系统的数据采集模块，并针对性的采集了金融相关的数据。在采集数据的基础上加入了关键词抽取、事件等级评估的功能。最后是在系统中尝试使用了三种不同的预训练语言模型，并比较了这三种模型生成摘要的效果。

1.2 国内外研究现状

事件检测任务一直是自然语言处理（NLP）领域的研究热点，吸引了国内外众多学者的关注。在国内，研究重点通常放在社交媒体和新闻领域的事件追踪上。例如，上海财经大学的研究团队提出了一种创新的商业事件提取方法，主要用于从在线中文新闻中提取关键商业事件[2]。清华大学的研究人员通过 ERNIE[3]模型优化了新闻事件的自动追踪和摘要生成过程，极大提升了关键信息提取的效率和准确性。在国际上，事件检测任务的研究同样取得了显著的进展。例如，最新的研究包括使用深度学习模型进行广义视频异常事件检测（GVAED）[4]，系统地分类不同的方法并确定新的研究方向，特别是在视频监控领域。这强调了深度学习模型在处理复杂视觉信息进行事件检测时的能力。同时，研究人员正在深入研究社交媒体的动态，以实时提取和分析事件，展示了针对 Twitter 等快节奏多元化平台的专门方法的有效性[5]。这一研究领域展示了通过实时数据处理，在监控和响应公共安全事件方面先进分析技术的潜力。

虽然这些研究不断推动语言模型在事件检测领域的技术进步，但它们通常缺乏针对特定行业的深度优化和针对性应用。通过深入分析金融行业的特定需求和挑战，本文旨在开发一个精准和高效的金融舆情事件检测系统，以提供对市场动态的实时监测

和分析。这不仅能够增强现有模型的应用范围，也为金融领域的决策支持和风险管理提供了新的技术工具。

1.3 研究目的与内容

1.3.1 研究目的

本研究旨在基于预训练的语言模型，开发和实现一个高效、准确的金融舆情事件检测系统。该系统的主要目的包括但不限于以下几点：

1.3.1.1 **提高检测效率和准确性：**利用预训练的语言模型的强大能力，提高对金融舆情事件的检测效率和准确性。相较于传统的基于规则或简单机器学习方法，预训练的语言模型不仅能够处理和分析复杂的文本数据，还可以捕捉细微的情感倾向和意图，从而在舆情事件的检测中发挥关键作用。此外，语言模型还能利用其自然语言处理能力，进行高级功能如情感分析、实体识别和话题聚类。这些高级分析能力可以进一步提瞭金融舆情监测的精度和深度，帮助金融机构和分析师更快速、更准确地识别市场动态和潜在风险。例如，通过话题聚类能够将相关的讨论和信息组织起来，为决策提供有力支持。

结合生成式模型的文本摘要和总结功能，还可以将大量的舆情数据压缩成易于理解的报告和概要，进一步增强决策的效率和质量。因此，利用预训练的语言模型不仅可以提高金融舆情事件检测的效率和准确性，还能在金融市场的监测和分析中提供全面和深入的视角。

1.3.1.2 **处理海量数据、增强实时监测能力：**在互联网和社交媒体时代，金融相关信息的数量庞大且持续增长。本系统旨在有效处理和分析这些大规模数据集，以识别出具有市场影响力的舆情事件。在此基础上，开发一个能够实时监测和分析金融市场舆情的系统。通过持续跟踪和分析互联网和社交媒体上的金融信息，系统将能够即时识别和报告可能影响金融市场的重要事件。

1.3.1.3 **支持投资决策、维护市场稳定：**通过及时了解市场动态，投资者可以更好地评估市场风险和机会。同时，及时识别和解读可能引发市场波动的舆情事件，有助于相关监管机构采取措施预防和缓解潜在的市场动荡，从而维护市场的稳定性和秩序。

通过实现上述这些目的，本研究期望为金融市场的舆情监测和分析提供一个更为高效、准确和全面的技术解决方案，从而为投资者、金融机构及监管机构在复杂多变的金融市场中提供有力的决策支持。

1.3.2 主要研究内容

1.3.2.1 数据收集

本研究的首要任务是收集适用于金融舆情事件检测系统的数据。主要目标是实现从诸如 Twitter 这样的社交媒体平台上获取金融相关的舆情事件的推文、帖子等文本数据。收集到的数据将经过预处理工作，包括按照时间切片、清洗等，为后续的数据处理和分析准备高质量的数据集。

1.3.2.2 推文分析

在采集到数据后，为了方便后续在系统中进行突发事件的检测和事件的摘要生成，需要在预训练的基础上对数据进行补充。这个部分主要是优化了一个基于知识抽取的文本关键词算法，以及在抽取出推文关键词的基础上人工制定了一个评估事件突发等级的算法。

1.3.2.3 比较预训练语言模型

在数据集准备完成后，本研究将着手比较不同预训练语言模型在金融舆情事件检测任务上的表现。计划比较的模型包括 PEGASUS-newsroom[6]、T5[7]、BART-Large-CNN[8]等。将从以下几个方面对这些模型进行比较：是否能高效处理提供的大量数据、是否能准确分析理解文本的内容、以及是否能有效的根据原始文本生成事件摘要、事件总结等。此环节的目标是从上述预训练的语言模型中筛选出在金融舆情事件检测任务上表现最优的模型，作为系统的核心算法。

1.4 论文组织

本文分为五个章节，每章的主要内容如下所示：

第一章：绪论。本章介绍了金融舆情事件检测技术及系统的研究背景、意义、本论文的研究基础、该领域的国内外研究现状以及本论文的主要研究内容。本章的最后介绍了本论文的组织结构。

第二章：系统整体介绍。本章对本研究改进并完善的金融舆情事件检测系统进行了整体的介绍。首先对系统的架构进行了整体的介绍，之后分别简要的介绍了前端和后端的功能模块。

第三章：数据采集与推文分析。本章详细介绍了本研究开发的用于从 Twitter 上获取指定关键词的推文数据的数据采集模块。该部分还介绍了数据采集模块除了获取数据之外的功能，包括数据预处理、推文关键词抽取和推文突发等级评估。

第四章：基于 BART-Large-CNN 实现金融舆情事件检测。本章首先从模型结构特点、适用任务领域介绍了 GPT[9]、BERT[10]和 BART 这三种常见的预训练语言模型，并结合本研究任务的要求，详细介绍了 BART-Large-CNN、PEGASUS-newsroom 和 T5 模型。本章的最后详细介绍了本研究构建的系统是如何使用语言模型来完成预定的功能的。

第五章：总结和展望。这部分简要概括了本文的研究内容，同时阐述了本研究的一些不足之处以及未来可以改进的方向。

2、系统整体介绍

2.1 系统架构介绍

本研究旨在使用 EDAS 系统完成金融舆情事件检测及事件分析，并在此基础上使用预训练语言模型进行一定的改进。系统整体采用的是一个简单的 B/S(Browser/Server)架构，在本地终端上提供一个用户友好的网页界面。系统主要分为前端展示交互层和后端处理分析层两个主要部分。系统整体架构如图 2-1 所示。

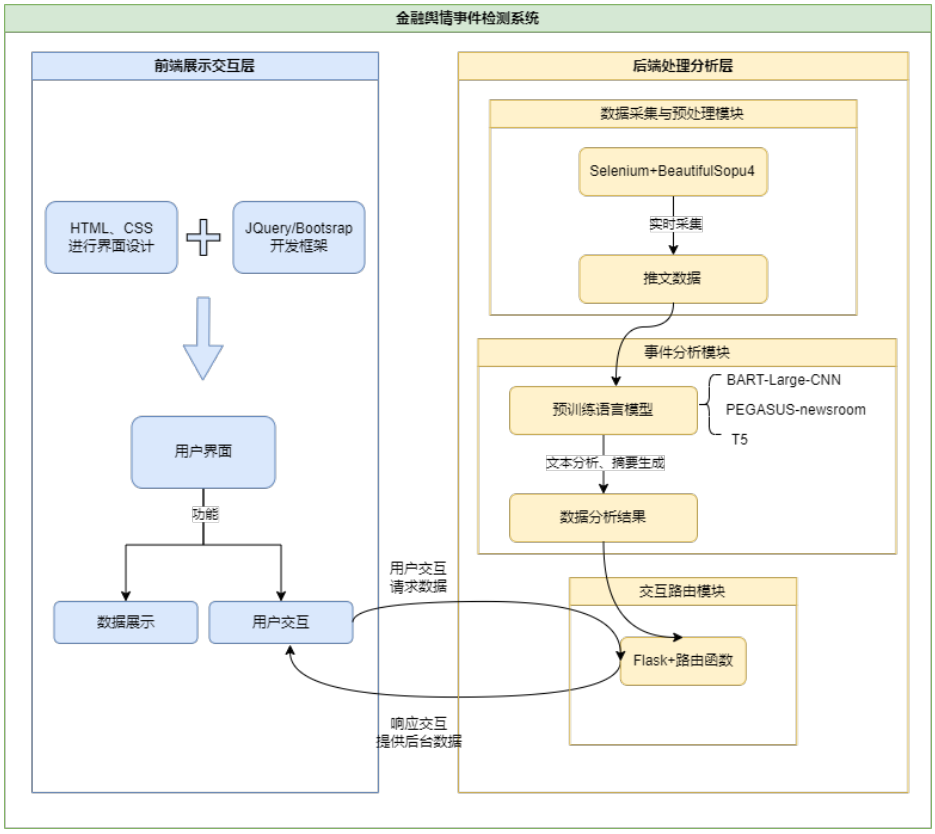

图 2-1 系统整体架构

Figure 2-1 Architecture of the System

2.2 前端介绍

前端部分基于 HTML 和 CSS 设计优雅、简介的网页界面，并使用 JQuery/Bootsrap 的开发框架以丰富交互、增强用户体验。这一层主要负责提供与后端的交互接口，使用户能够上传数据、选择数据范围和执行关键词检索等操作。此外，前端还负责将数据分析的结果以表格、折线图、堆叠图、词云和知识图谱等多样化和直观的形式展现给用户，以便用户快速获取所需信息。前端由两个页面组成，分别是实时监测页面和事件分析页面。

实时监测页面允许用户上传需要进行趋势检测的数据集。经过数据处理和分析，该页面将展示数据集内各事件的时间、推文内容（突出关键词）及事件等级。为了让

用户更直观地理解数据分析结果，页面通过折线图、词云，以及在世界地图上标注突发事件发生地等多种形式来呈现信息。这种设计使得实时监测既全面又直观，有助于用户迅速捕捉和理解正在发生的趋势和事件。

事件分析页面根据用户指定的时间和事件关键词，展示所有相关事件的统计和分析结果。这个页面设计让用户能直观地识别出哪些日期发生了重大突发事件，并提供相关事件的摘要及详细信息。时间线展示模块在时间线上标注了重大事件发生的日期，同时列出了相关事件的摘要，为用户提供丰富的信息资源。这种布局和功能设计，使得用户能够有效地追踪和分析时间序列中的重要事件，增强了事件洞察的深度和广度。

2.3 后端介绍

后端部分构成了系统的数据处理和分析核心，它由数据采集与预处理模块、事件分析模块和交互路由模块组成。数据采集与预处理模块利用 Selenium 和 BeautifulSoup4 自动化地从 Twitter 抓取含指定关键词的推文，并完成数据的清洗和时间切分。之后，数据传递至事件分析模块，这是系统的关键组成部分，核心依托于预训练的语言模型，在本研究中分别使用了 BART-Large-CNN、PEGASUS-newsroom 和 T5 这三个模型进行尝试。该模块通过利用语言模型在文本分析和摘要生成方面的能力，执行关键词提取、突发事件检测及等级评估、事件摘要生成等多项任务，从而完成对数据的深入分析。分析完毕的数据将被储存于本地终端。当用户通过前端界面发起请求时，交互路由模块会根据需求选取相应的文件和数据，并经过精细化的可视化处理后在前端展示。

2.3.1 数据采集与推文分析模块

该模块主要实现数据的采集功能。基于对 Twitter 反爬虫机制的了解和对其网页结构的研究，最终决定通过 Selenium 库来模拟网页上的人工操作（如搜索、滚动、点击），并通过 BeautifulSoup4 解析 Twitter 网页的 HTML 来获取需要的推文数据。获得数据后，该模块还会对采集到的推文进行一些分析，提取出推文中的关键词并根据关键词完成事件突发等级的评估。

2.3.2 交互路由模块

交互路由模块基于 Python 的 Flask 库构建。交互路由用于处理前端请求，允许用户与后端系统进行有效的通信。利用 Flask 的灵活性和易用性，该模块可以快速响应各种 HTTP 请求，如 GET、POST 和 PUT。这些请求通常涉及从前端收集用户输入、处理这些数据，并返回适当的响应或结果。

2.3.2 事件分析模块

事件分析模块是本系统的核心模块。预计分别使用 Meta AI 发布的 BART-Large-CNN、Google 发布的 PEGASUS-newsroom 和 T5 这三种语言模型作为事件分析模块的核心功能部分，并对这三个模型的摘要生成效果进行一定的比较。

2.4 本章小结

本章对本研究完善并改进的金融舆情事件检测系统进行了整体的介绍。系统的需求分析部分强调了实时分析金融文本数据的重要性，以及系统在识别重大金融市场事件和对其市场影响进行预测的能力。采用 JQuery/Bootstrap 与 Flask 框架，实现前后端分离的开发策略，确保了系统的兼容性和实时数据处理能力。

之后，详细介绍了系统前后端的设计思路。前端设计部分着重于用户界面，包含实时监测、事件分析、事件图谱等子页面，通过图表、词云和地图等多种形式直观展示信息，使得用户能够快速理解和追踪金融事件。后端设计部分包括数据采集与处理模块、交互路由模块和事件分析模块。数据采集模块通过 Selenium 和 BeautifulSoup4 来克服爬虫限制，获取推文数据。交互路由模块定义了多个路由函数，支持文件上传、数据分析和结果展示。事件分析模块计划使用 Meta 的 BART-Large-CNN、Google 发布的 PEGASUS-newsroom 和 T5 这三种语言模型三种语言模型来优化文本摘要生成。

总体而言，该系统通过整合前端交互设计和后端数据处理，旨在提供一个高效、易用的金融舆情监测平台。

3、数据采集与推文分析

数据是一个系统的基础，而原本的系统中并没有有效的数据采集模块。因此，需要针对 Twitter 的网页格式和数据情况设计一个实用、高效的数据采集模块，并对数据进行一定程度的预处理。在完成数据预处理的基础上，会对推文文本进行基础的分析，抽取出文本的关键词，并基于关键词完成事件突发等级的评估。完整的数据采集与推文分析的流程如图 3-1 所示。

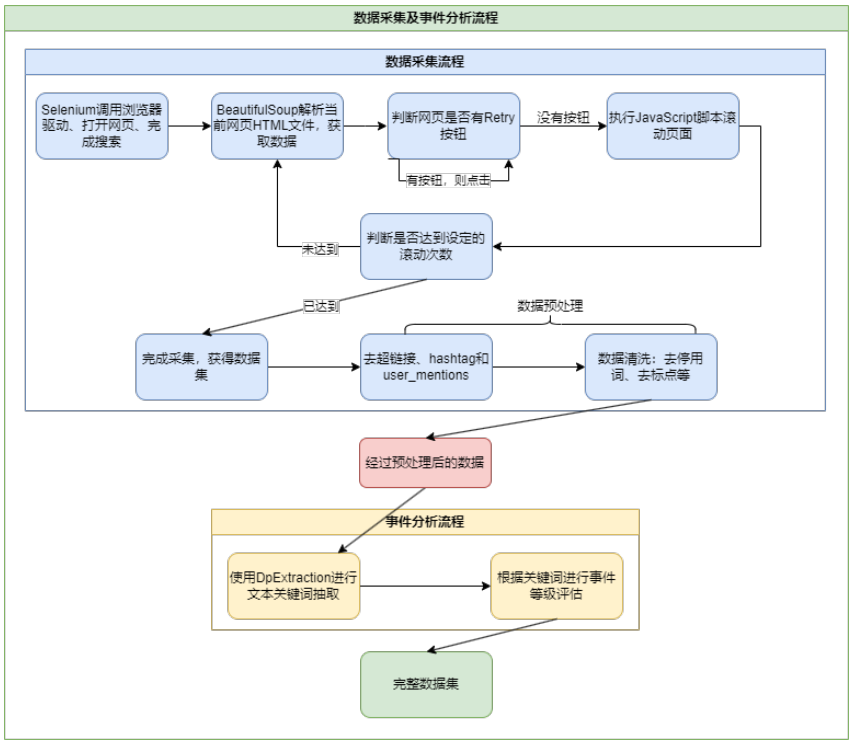

图 3-1 数据采集与推文分析流程

Figure 3-1 Data Acquisition and Tweet Analysis Process

数据采集并完成推文分析后的格式如图 3-2 所示。

```
tweet_contents.append({
  "created_at": formatted_time,
  "id": tweet_id,
  "text": cleaned_text,
  "retweet_count": retweet_count,
  "entities": {
      "hashtags": hashtags,
      "user_mentions": mentions,
  },
  "origin_text": content_text,
  "keywords": keywords,
  "level": level
})
```

图 3-2 数据格式

Figure 3-2 Data Format

3.1 数据采集模块

Twitter 作为全世界最大的公共社交媒体平台之一，聚集了来自世界各地的、大量的用户的推文，其中有很多与时事密切相关，因此 Twitter 上的推文是很有价值的数据来源。但 Twitter 为了保护用户的隐私安全，设置了严格的反爬虫机制。比如，必须登录 Twitter 账号才能正常访问推文、限制单名用户每条能访问的推文的数目（5000 条左右）以及限制用户访问推文的速率等。并且，由于 Twitter 的网页结构发生了比较大的变化，如 Github 这样的开源平台上的相关的爬虫代码已经无法使用。因此，需要根据对数据的需求，重新学习 Twitter 的网页结构以实现数据采集模块。

根据上文对 Twitter 的反爬虫机制的研究和对 Twitter 网页结构的学习，最终决定通过 Selenium 和 BeautifulSoup4 库实现数据采集模块。这两个库分别用于模拟网页操作和 HTML 内容解析。Selenium 允许程序模拟人工操作，如打开网页、输入查询条件、滚动页面等，而 BeautifulSoup4 用于解析 HTML 代码，提取和处理网页中的数据。本模块的主要目标是自动化地从 Twitter 上采集关于特定话题（例如'#Finance'）的推文数据，以便进行后续的金融舆情分析。

具体来说，数据采集模块首先使用 Selenium 调用 ChromeDriver 浏览器驱动，从而启动 Chrome 浏览器并访问 Twitter 网页。为了应对 Twitter 必须登录才可以正常访问推文的限制，确保可以访问和抓取数据，模块会加载预先保存的用户 cookies，这样可以模拟已登录用户的会话，从而绕过登录验证获取更丰富的内容。

由于推特页面上的内容是动态加载的，需要不断滚动页面才能持续加载出新的推文数据。因此，该模块通过执行 Javascript 脚本 window.scrollTo(0, document.body.scrollHeight)实现了自动滚动功能，模拟用户滚动浏览器窗口以加载更多的推文。模块通过设置一定的滚动次数（一般是 1000 次到 5000 次之间），来尽可能多的获取数据。另外，通过设置适当的两次滚动之间的暂停时间，可以模拟人类浏览行为，有效避免因操作过快而触发 Twitter 的反爬虫机制。为了更贴近人类的操作，滚动暂停时间在 3.8 秒到 4.2 秒之间通过随机数函数生成。

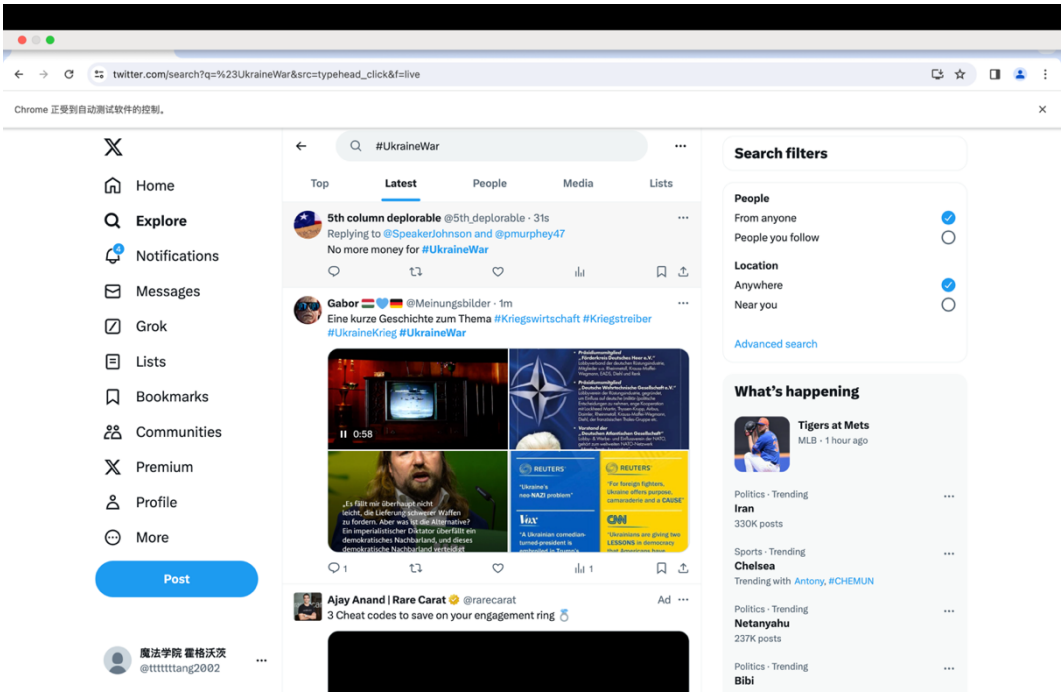

图 3-3 爬虫代码运行示例

Figure 3-3 Web Crawler Code Execution Example

遗憾的是，尽管已经设计了随机的滚动暂停时间来模拟人工操作，在程序实际运行过程中，还是发现会出现访问推文被限制的情况。具体来说，当滚动次数达到一定数量后，滚动加载新推文会被限制，页面上会出现 Retry 按钮并且出现访问被限制的提醒，需要不断点击页面上的 Retry 按钮才能继续获得新数据。若是没有点击该按钮，程序仍然会继续运行，滚动次数仍然在增加，但页面实际一直停留在当前页面，没有获取到新数据。针对这一问题，数据采集模块设计了在每一次滚动之后，通过直接查找 Retry 按钮对应的按钮元素来判断页面上是否出现 Retry 按钮，如果按钮存在则持续点击该按钮并继续查找、直到在页面上找不到 Retry 按钮。这样的方法能够基本避免 Twitter 限制访问导致无法访问新数据的问题，但这种方法还不完善——程序运行过程中发现，不断点击 Retry 按钮可能导致按钮的位置发生偏移，当按钮被遮挡时（例如被浏览器顶栏遮挡），会导致判断结果为没有 Retry 按钮，继而再次出现滚动次数增加，但网页实际并没有刷新、没有获得新数据的情况。为了解决这个问题，改用直接查找 Retry 按钮对应的 XPATH、通过元素路径精确定位页面上的按钮的方法，最终成功实现了只要有 Retry 按钮、无论是否被遮挡，都能正确找到按钮并点击的功能，从而完全解决了访问被限制导致的无法获取新数据的问题，真正实现了持续的获取数据。但 Twitter 对用户访问速率的限制仍然无法避免，因此程序获取数据的速度比较慢，平均滚动 1000 次需要 1 小时，每 1000 次滚动平均获得 3000 条数据。

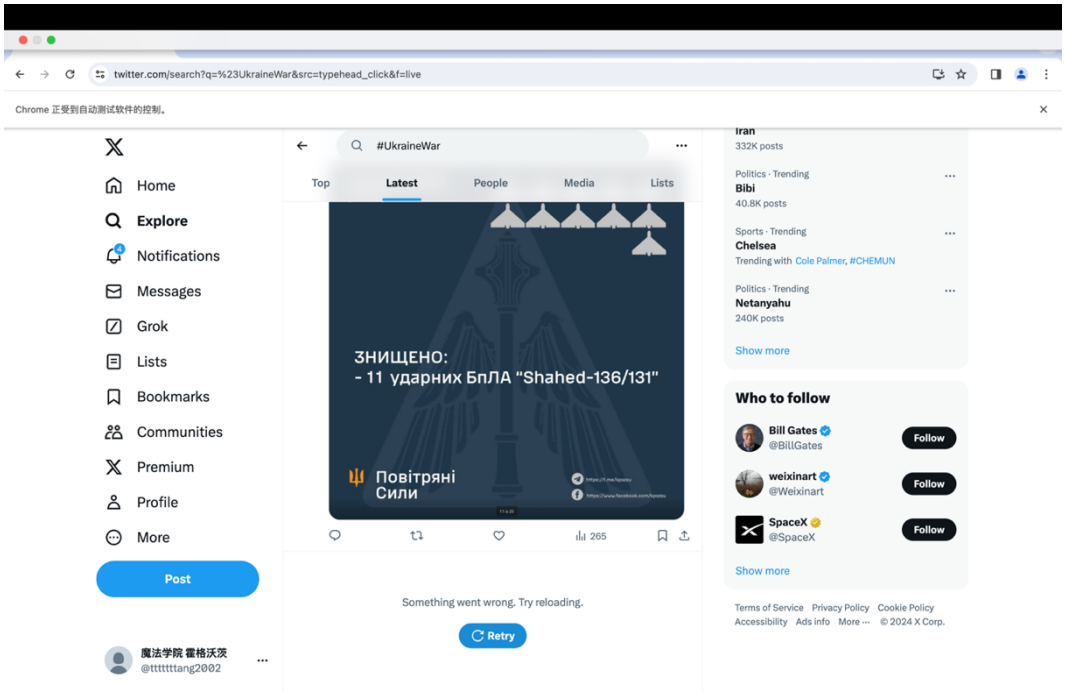

图 3-4 推文访问被限制示例

Figure 3-4 Tweet Access Restricted Example

当新的推文数据加载完成后，数据采集模块使用 BeautifulSoup4 解析当前页面的 HTML 文件，根据需要的数据在网页中对应的元素提取出每条推文的详细信息，包括推文原始文本、发布时间、语言、推文 ID、转发次数、以及推文中包含的 hashtags 和 user_mentions（也就是@的人名）。这一过程涉及对 Web 元素的细致选择和数据的严格筛选，以确保获取的信息准确无误。此外，为了便于后续的分析，在数据采集时便会同步完成推文文本的预处理。预处理的主要步骤便是去除文本中的超链接、hashtag 和提到的人名，同时还会完成常规的去停用词、去标点等数据清洗过程，最终只保留纯粹的文本。清洗后的文本会储存在数据的“text”字段里。

经过在实践中不断发现问题、解决问题，本研究成功开发了一个可以稳定运行的 Twitter 数据采集模块，用来完成 Twitter 数据的采集。本研究后续用到的所有数据都是由这个数据采集模块获取的。本研究根据“#Finance”这个关键词采集了推文数据，来分析系统对舆情事件、特别是金融领域的舆情事件检测和摘要生成能力。

具体来说，数据都是在 Twitter 上检索选择好的 hashtag、并从搜索结果的 Latest 页面获取到的。从 Latest 页面获取数据的好处是，这个页面按照推文的创建时间从最新到最旧排序，并且该页面收集了所有包含搜索的 hashtag 的推文，从而保证了数据的完整性、并且省去了后续需要对推文按照时间进行排序的步骤。总计获取到了从 2024 年 2 月 18 日到 2024 年 4 月 10 日的、包含“#Finance”的 49535 条数据。

表 3-1 数据集介绍

Table 3-1 Introduction of Dataset

数据类别	“#Finance”
数据总数目	49535
推文时间范围	2024-02-18 到 2024-04-10
推文平均长度	28 个单词
平均 hashtag 个数	7 个 hashtag

3.2 推文分析模块

如图 3-5，在获取到推文数据后，将会基于依存句法分析和深度学习结合的知识抽取来完成文本的关键词抽取，并在抽取出关键词后完成推文突发等级的评估。

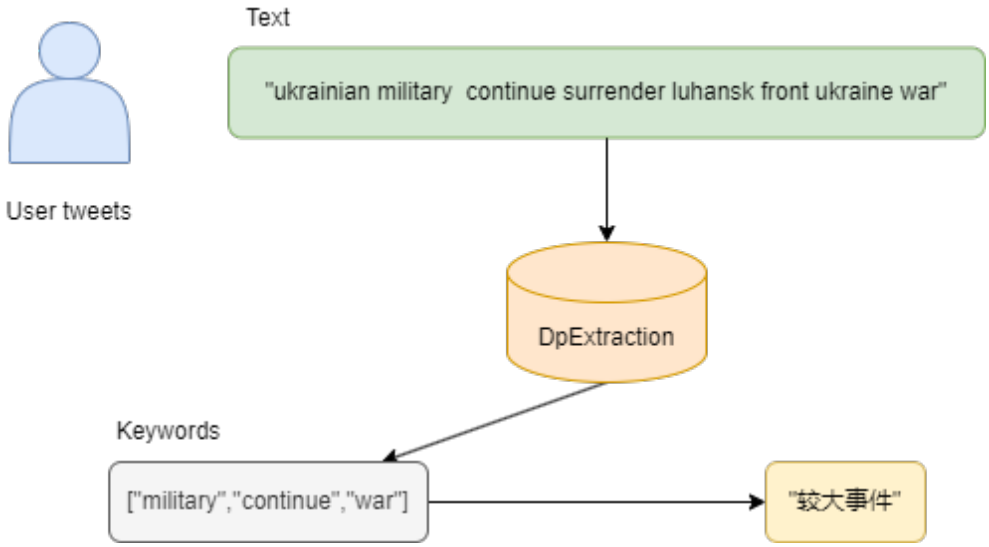

图 3-5 推文关键词抽取、等级评估示例

Figure 3-5 Tweet Keyword Extraction and Rank Assessment Example

具体来说，该模块定义了一个“DpExtraction”类来完成相关工作。“DpExtraction”类是一个使用 Stanford CoreNLP 工具[11]进行句法分析（dependency parse）和信息抽取的工具类。它包含了一系列方法，用于从句子中提取主谓宾、主系表、被动语态等语法结构，并将这些结构组织成三元组的形式。通过调用这个类中定义的 get_triples 函数，可以提取到句子中的主语、谓语和宾语，作为一条推文的关键词。具体来说，输入的句子首先会被分词，得到句子中词与词的依赖关系。之后，根据词语之间的依赖关系，提取出句子中的主谓宾（Subject Predicate Object, SPO）、主系表（Subject

Predicate Complement, SPC）和被动语态（Passive voice）。最终只保留抽取出的主谓宾结构作为原始推文的关键词。

但这样的关系词抽取方法存在一定问题。在进行关键词抽取时，使用的是预处理后的文本，但因为有的推文文本的关键信息（例如主、谓语）在 hashtag 或者 user__mention 中，而清洗时会将这些信息去除，从而导致无法提取出句子中的主谓宾结构。为了完善关键词抽取算法，设置当 get_triples 函数返回值为空时（代表没有抽取出推文的主谓宾关系），从该条推文的 hashtags 和 user_mentions 中选择词语作为关键词。为了选取出最具代表意义的 hashtags 和 user_mentions，首先需要统计所有 hashtags 和 user_mentions 的出现次数，并在无法抽取出推文的主谓宾关系时选择该条推文出现最多两个的 hashtag 和一个 user_mentions，作为该推文的关键词。如果 hashtags 和 user_mentions 的总数不足三个，则按照出现次数的排序进行重复。

最后，根据每条推文的关键词出现频率，将所有推文划分为四个等级，分别是“特别重大事件”，“重大事件”，“较大事件”和“一般事件”。推文突发等级评估的准则是：先统计所有关键词出现的次数，并记录下最高的出现次数。再分别计算每一条数据的三个 keywords 出现次数之和，如果总出现次数大于最大出现次数的 90%，则判定该事件为“特别重大事件”；小于等于 90%但大于 60%，则判定为“重大事件”；小于等于 60%但大于 20%为“较大事件”，否则为“一般事件”。在这个规则下，对上文介绍的本系统获取的“#Finance”数据集进行推文突发等级评估结果如表 3-2 所示：

表 3-2 推文突发等级评估结果 1

Table 3-2 Tweet Severity Assessment Results 1

事件等级	事件数目
'特别重大事件'	12627
'重大事件'	3998
'较大事件'	2584
'一般事件'	30326

上表中“特别重大事件”的数量明显多于“重大事件”和“一般事件”，经过思考，这应该是评估准则不合理导致的问题。例如，一旦某条推文中包含出现次数最多的 keyword，那这条推文必然被认为是“特别重大事件”，而包含出现次数最多的 keyword 的推文也是最多的，这就会导致“特别重大事件”数目偏多，这显然是有失

准确的。因此，需要改进评估逻辑，计算一个事件关键词的平均出现频率而不是总和，这样可以相对更公平地评估每个事件的突发程度。

表 3-3 推文突发等级评估结果 2

Table 3-3 Tweet Severity Assessment Results 2

事件等级	事件数目
'特别重大事件'	179
'重大事件'	491
'较大事件'	15955
'一般事件'	32910

改进评估逻辑后，推文突发等级评估结果如表 3-3 所示。可以看出，此时的事件等级分布不再有之前的“特别重大事件”过多的异常，是比较符合逻辑的评估准则。

采集和处理完的数据最终被保存为 JSON 格式文件，以便进行后续的分析和处理，确保金融舆情事件检测系统能够高效利用最新和相关的数据，进行后续的突发事件检测和事件摘要生成。

3.3 本章小结

本章节介绍了针对 Twitter 推文的数据采集与分析方法。首先，由于 Twitter 对数据采集有严格的限制，如需登录访问、单个用户可访问的推文数量限制及速率限制等，因此开发了一个高效的数据采集模块。这一模块主要通过 Selenium 模拟浏览器操作和 BeautifulSoup4 解析 HTML，从而绕过登录验证，实现对特定话题推文的自动化采集。尽管存在访问限制，通过模拟用户滚动操作和处理“Retry”按钮，模块能够持续采集数据。

数据采集后，进行预处理，包括去除超链接、hashtag、提到的人名等，保留纯净的文本用于分析。分析模块利用 Stanford CoreNLP 工具进行依存句法分析，从推文中抽取关键词并基于抽取出的关键词人工制定了推文突发等级评估规则。最初的评估方规则存在偏差，主要体现在对“特别重大事件”的判定过多。后续，通过调整关键词出现频率的计算方法，得到了更合理的事件等级分布。

综上，本章详细介绍了针对 Twitter 推文的数据采集方法，并在采集到数据的基础上介绍了数据预处理和基础的推文分析的方法。

4、基于 BART-Large-CNN 实现金融舆情事件检测

4.1 模型介绍：BART-Large-CNN

4.1.1 BART 模型概述

BART（Bidirectional and Auto-Regressive Transformer）模型是由 Facebook AI 在 2019 年推出的先进语言处理工具，它巧妙地结合了编码器-解码器架构，吸收了 BERT 和 GPT 模型的优势。其编码器部分借鉴了 BERT 的双向 Transformer[12]机制，这使得 BART 能够全面地理解输入文本的上下文，而其自回归解码器则模仿了 GPT 的生成方式，逐个词汇地预测文本，从而在文本生成任务中尤为高效。BART 的训练过程包含了一系列去噪策略，如文本缺失填充和句子重排，通过这些预训练任务，模型能在处理不完整或混乱的文本时重构和细化原始信息。

如图 4-1 展示了 BART 模型的结构。与纯编码器模型 BERT 相比，BART 通过其解码器部分增强了文本的连贯生成能力，适应范围更广；与解码器模型 GPT 相比，它通过双向编码器更全面地把握语境和内容细节。这种结构的设计让 BART 特别适合处理如事件检测和分析等任务，因为它不仅能够从大量文本中准确地识别关键信息，还能结合上下文生成连贯、详细的事件描述。在事件检测中，BART 可以通过深入的语义理解来识别事件的发生和类型；在事件分析中，它则能够综合各类信息，生成有深度的分析报告。因此，BART 模型在文本理解和生成的多样性、深度以及准确性方面显示出明显的优势，尤其在需要高级理解和生成能力的应用场景中，如新闻摘要、内容创作、自动翻译及问答系统，都能体现其出色的性能。

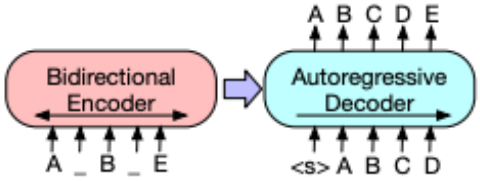

图 4-1 BART 模型结构

Figure 4-1 BART Structure

4.1.2 预训练语言模型的功能和重要性

预训练语言模型（PLMs）如 BART-Large-CNN 已成为自然语言处理（NLP）领域的核心，因为它们能够捕获大规模文本数据中的语言规律和语义结构。这些模型通过在广泛的文本上进行预训练，学习到复杂的语言特征，然后可以通过微调适应特定的 NLP 任务，如文本摘要、情感分析和事件检测。预训练语言模型的核心优势在于

其通用性和灵活性：一旦预训练完成，相同的模型可以用于多种任务，大大减少了从头开始训练模型的需要。

在金融舆情事件检测中，预训练模型的重要性体现在它们对复杂文本的深度理解能力。金融文本常常包含专业术语和隐含信息，PLMs 通过捕获这些细节来辅助分析师理解市场动态和预测趋势。例如，BART-Large-CNN 通过理解文本的上下文和细节，可以帮助检测金融新闻或社交媒体中的突发事件，并生成摘要，为快速决策提供支持。

4.1.3 BART-Large-CNN 模型介绍

由 Facebook AI（现在称为 Meta AI）开发的 BART-Large-CNN 是一种专门为文本摘要优化的高级语言处理模型，特别适用于新闻领域。这一模型基于 BART 架构，整合了强大的双向编码器和自回归解码器。双向编码器通过考虑文本中每个词语的上下文信息来增强语义理解，而自回归解码器则逐字生成连贯的摘要文本，确保输出的逻辑性和连贯性。

在 BART-Large-CNN 中，注意力机制是实现深层语义理解和生成任务的关键。BART 模型结合了双向编码器的全面理解能力和自回归解码器的生成能力，使其在文本处理任务中表现卓越。多头注意力机制使 BART 能够在编码阶段捕获文本的不同方面，而在解码阶段生成连贯、相关的输出。

BART-Large-CNN 的预训练过程涉及在大量新闻相关数据上的深度学习，这一过程训练模型捕捉语言细节和复杂的叙事结构，使其特别擅长生成精确且信息丰富的摘要。此外，模型的“CNN”名称不仅代表其神经网络结构，还强调了在新闻内容摘要生成中的特定优化和卓越表现。BART-Large-CNN 之所以优越，是因为它精细地平衡了深入的文本理解和高效的生成能力，使其能够快速准确地从大量新闻报道中提取关键信息，并转化为高质量的摘要。

在本系统中，BART-Large-CNN 利用这种深度理解生成突发事件的摘要。它首先通过编码器理解整个输入文本，包括金融市场的细节和复杂情境，然后解码器基于这种理解生成准确的事件摘要。这种基于深度学习的方法能够处理大量数据，快速识别和总结金融市场的关键变化，对金融专业人士和分析师提供有价值的洞察。

4.2 用 BART-Large-CNN 进行实验

预训练语言模型是本研究实现的系统的核心部分，依赖于生成式语言模型强大的文本分析能力和摘要生成能力，可以对经过预处理后的推文文本进行突发事件检测、事件摘要生成等。为了探究预训练语言模型对舆情事件、特别是金融舆情事件进行检

测和事件摘要的效果，在本系统中使用 BART-Large-CNN 模型完成上述任务，并使用 PEGASUS-newsroom 和 T5 进行对比实验。

本研究使用第三章得到的、经过了预处理和初始分析的数据的基础上进行突发事件检测，筛选得到价值较高的突发事件聚类，之后再使用预训练语言模型对这些事件进行文本分析和事件摘要生成，整体分析流程如下图 4-2 所示。

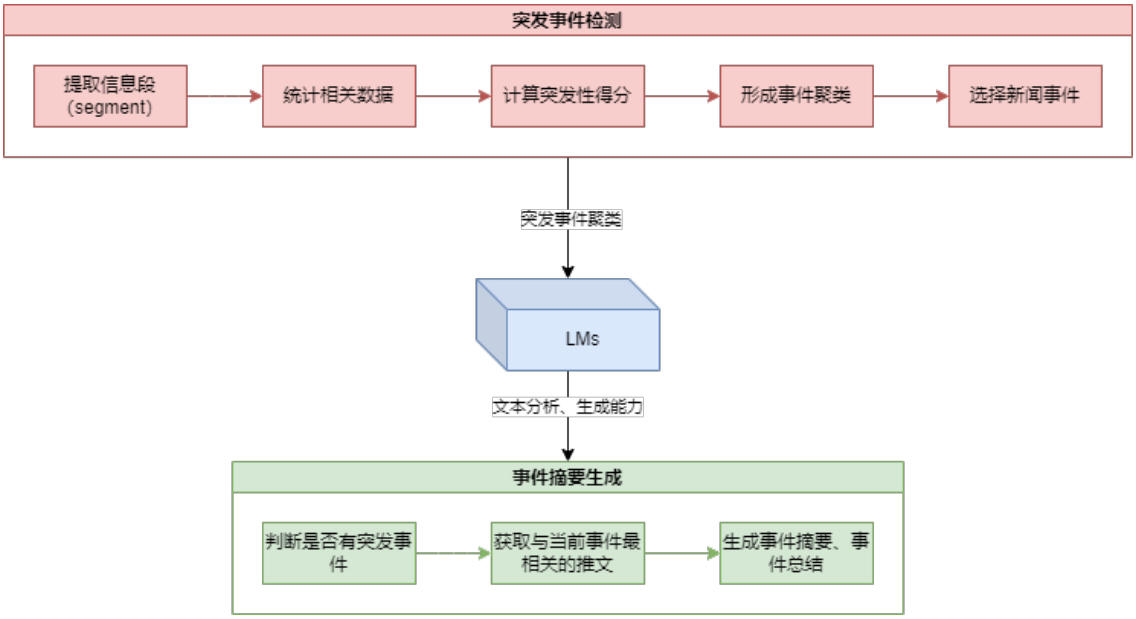

图 4-2 数据分析流程

Figure 4-2 Data Analysis Process

4.2.1 突发事件检测

进行突发事件检测时，会遍历数据切分后存储的目录，并对每个时间段（也就是每个日期）分别进行如下处理：

1). 提取 segment：将推文文本的 token（也就是推文的每一个单词）、推文的 hashtags 和推文中提到的人名尽可能的与维基百科页面的标题进行匹配，从而提取出有潜在的突发性和新闻价值的文本片段，称为 segment。

2). 统计相关数据：统计每个 segment 的在该段时间内出现的频率、涉及每个 segment 的用户数（也就是发的推文中包含该 segment 的用户的数目）以及相关的推文的转推总数。

3). 计算突发性得分：根据上述信息，利用 sigmoid 函数和对数转换来计算每个 segment 的突发性得分，以强调 segment 的重要特征和差异。

4). 形成事件聚类：基于 segment 的相似度形成事件聚类，每个聚类代表一个可能的事件，其中的节点是相关的 segment。

5). 选择有价值的新闻事件：基于事件中的 segment 的得分和相似度计算每个事件的新闻价值，再用新闻价值对事件进行筛选，只有得分高于设定的阈值的事件才会被保留。最终会得到一组满足阈值要求的事件，每个事件由一组信息段和相应的突发性得分组成。

4.2.2 事件摘要生成

完成突发事件检测之后，便会开始对突发事件进行摘要，这也是本系统最关键的功能之一。为了高效、准确的生成突发事件的摘要和总结，便需要依靠上文提到的生成式语言模型的文本分析和摘要能力。生成摘要的过程如下：

1). 判断当前时间段是否有突发事件发生：在事件检测阶段，形成事件聚类之后，会保留满足阈值要求的事件。如果当前事件段所有事件都不满足阈值要求，则代表没有值得关注的突发事件发生，不需要进行之后的事件摘要等操作。

2). 获取与当前事件最相关的推文：通过考虑与当前事件相关的推文中的 segment、推文结构和关键词，筛选出与该事件最相关的 5 条推文，并将每条推文的权重和推文文本写入文件中，形成该事件的摘要。

3). 生成事件总结：这一步使用 transformer 库中的 pipeline 函数[13]完成预训练语言模型的调用。将上一步选出的推文整合后作为输入传给预训练语言模型，并将 pipeline 函数的“task”参数设置为“summarization”，意为下达让预训练语言模型完成事件总结任务的指令，从而有效的生成当前事件的总结。

4.2.3 实验结果

通过以上步骤，得到了如图 4-3 的事件总结，它包含了突发性文本片段列表、事件突发性判断以及事件摘要。其中突发性文本片段列表给出了突发性文本片段及其得分，并按得分从高到低排序。

图 4-3 数据处理结果

Figure 4-3 Data Processing Results

4.3 用 PEGASUS-newsroom 和 T5 进行对比实验

BART-Large-CNN 是一个先进的基于 Transformer 架构的序列到序列模型，它通过结合预训练和微调策略，优化了文本生成任务，尤其在新闻摘要领域表现卓越。为了深入探究 BART-Large-CNN 在处理舆情事件、特别是金融领域舆情事件的摘要生成效果，本研究计划与其他顶尖模型进行对比实验。选取的对比模型包括 PEGASUS-newsroom 和 T5，这两者都基于 Transformer 架构，且在文本摘要任务上具有杰出表现。这三个模型在预训练方法上有所区别：T5 采用广泛任务的文本数据进行预训练，适用于多样的文本处理任务；PEGASUS-newsroom 则专注于新闻领域，通过预训练新闻摘要任务的特定文本数据来增强其对新闻文本的处理能力；BART-Large-CNN 则采用了混合策略，不仅预训练了特定新闻领域的文本，也覆盖了广泛任务的文本数据，从而在广泛性和专业性之间取得平衡。通过这样的对比实验可以更全面地评估 BART-Large-CNN 在金融舆情摘要生成中的表现，并探索各模型的优势和局限，为未来的应用提供科学的参考依据。

4.3.1 PEGASUS-newsroom 模型介绍

PEGASUS-newsroom（Pre-training with Extracted Gap-sentences for Abstractive Summarization）是基于 Google 的 PEGASUS 模型并专门针对新闻报道进行优化的语言模型。该模型通过其创新的 Gap Sentences Pre-training (GSP) 方法在预训练阶段选择信息密集的句子作为缺口句子，并用剩余的内容来预测这些句子。这种方法使模型在从大量文本中抽取核心信息以生成精炼、内容丰富的摘要方面表现出色。特别是在金融新闻领域，PEGASUS-newsroom 能够识别重要的市场动态和事件，快速生成新闻摘要，帮助金融分析师理解市场趋势、公司财报和行业动态。通过自动化处理大量新闻数据，它显著提高了金融分析的效率，为快速发展的金融市场和实时决策提供了支持。此外，PEGASUS-newsroom 可以根据具体的金融需求进行定制化训练，增强其在特定金融文本分析中的性能和灵活性，成为金融领域处理和分析新闻内容的强大工具。

4.3.2 T5 模型介绍

T5 模型（Text-to-Text Transfer Transformer）由 Google 研发，是一种通用的文本到文本转换模型，旨在处理各种自然语言处理任务。T5 将所有任务统一视为文本转换问题，无论是文本摘要、翻译、问答还是分类任务。它通过大规模的预训练，学习了大量的语言知识和文本结构，这让它能够理解并生成复杂的文本内容。在金融领域，

T5 模型特别适用于解析金融报告、生成投资建议或进行市场趋势分析。它的强大之处在于能够处理大量复杂的金融数据和文本，通过灵活的文本到文本框架来提供深入的分析和洞察。此外，T5 可以根据特定的任务需求进行微调，使其更好地适应金融领域的特定语境和需求。因此，T5 模型不仅提高了金融数据处理的效率和精确度，还为金融决策支持提供了强有力的分析工具。

4.3.3 模型摘要对比

下表 4-1 展示了三个模型对于采集到的推文中的部分重大事件生成的摘要。

表 4-1 模型摘要对比

Table 4-1 Model Summary Comparison

事件描述	BART 摘要	PEGASUS 摘要	T5 摘要
Tata Power Solar的重大项目	Tata Power Solar completes India's largest Solar & Battery Energy Storage Project, a game-changer in the country's renewable energy landscape.	Browse Tata Power Solar latest news and updates, watch videos and view all photos and more. Join the discussion and find more about Tata Power Solar at abcnews.com	Tata Power Solar completes India's largest solar & battery energy storage project . project is a game-changer in the country's renewable energy landscape .
L&T 与 NVIDIA 的合作	L&T Technology Services partners with NVIDIA to train 1,000 engineers in generative AI, driving innovation across industries.	Browse NVIDIA latest news and updates, watch videos and view all photos and more. Join the discussion and find more about NVIDIA at abcnews.com	L&T Technology Services partners with NVIDIA to train 1,000 engineers in generative AI, driving innovation across industries .

续表 4-1 模型摘要对比

Continuation of Table 4-1 Model Summary Comparison

Axis Bank 信用卡欺诈事件	Axis Bank credit card holders affected by fraudulent international transactions should take immediate action to safeguard their finances. thanawala_hiral has all you need to know.	Axis Bank credit card holders affected by fraudulent international transactions should take immediate action to safeguard their finances. thanawala_hiral has all you need to know	credit card holders affected by fraudulent international transactions should take immediate action to safeguard their finances . thanawala_hiral has all you need to know https://moneycontrol.com/news/business/personal-finance/axi
ABT 的表现	$ABT is out performing your favorite basketball team. It's incredible to see how technology is democratizing investment opportunities. Looking forward to seeing where this trend goes!	Tokenization opens up access to commodities like never before! It's incredible to see how technology is democratizing investment opportunities. Looking forward to seeing where this trend goes! #Innovation #Finance #DigitalAssets $ABT is out performing	tokenization opens up access to commodities like never before . it's incredible to see how technology is democratizing investment opportunities . $ABT is out performing your favorite basketball team .

综合比较，三个模型各自从不同的角度展现了事件摘要的能力。BART-Large-CNN 以其内容的丰富度和详细性突出，表现出优异的信息完整性和相关性，非常适合需要深入信息和背景知识的场景。PEGASUS-newsroom 则倾向于包含外部链接和商业性

信息，这种风格使其在新闻报道和市场营销领域中特别有效，能够增强读者的参与感。另外，T5 在生成摘要时展现出高度的简洁性和直接性，适用于迅速获取事件核心信息的需求。综上所述，尽管在精确度方面仍有提升空间，BART-Large-CNN 全面和详尽的事件描述能力满足了本系统对推文内容进行深度分析、生成详尽的事件摘要的需求，因此本系统最终决定使用 BART-Large-CNN 模型。

4.4 实验成果展示

本系统最终的目的是将由推文分析模块、突发事件检测模块的分析结果和预训练语言模型生成的摘要在前端以表格、折线图、词云等多种方式进行直观的展示，从而使用户能直观的从中获取信息。下图 4-4、4-5 展示了最终前端页面的数据展示效果：

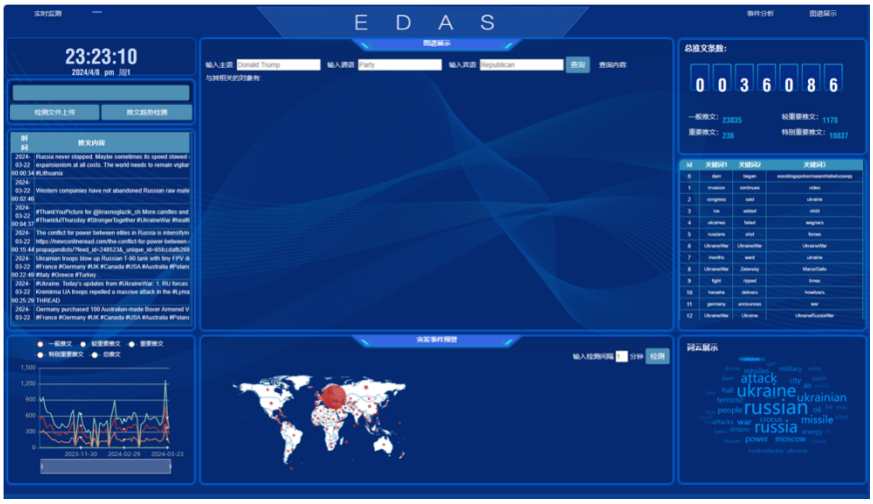

图 4-4 数据展示效果——实时监测页面

Figure 4-4 Data Display Effect — Real-time Monitoring Page

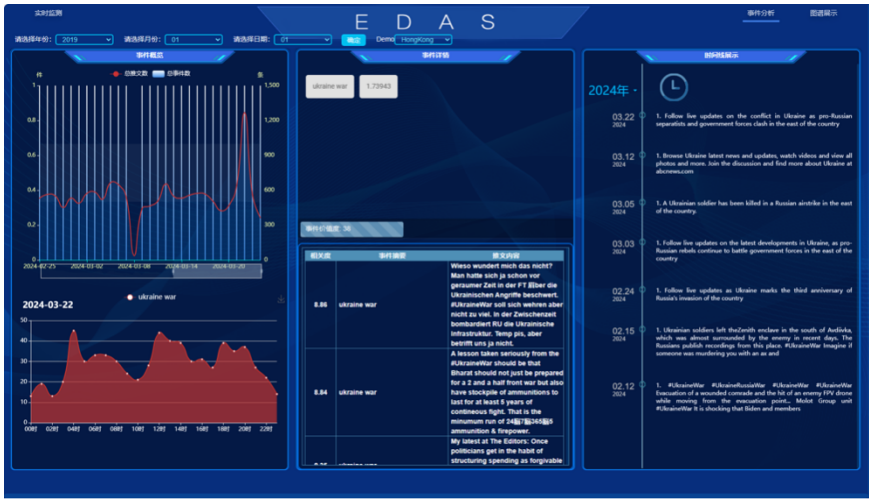

图 4-5 数据展示效果——事件分析页面

Figure 4-5 Data Display Effect — Event Analysis Page

4.5 本章小结

本章详细介绍了本研究如何基于 BART-Large-CNN 模型完成金融舆情事件检测任务。本章开头详细介绍了 BART 模型的结构和功能，并通过与 BERT 和 GPT 模型的对比得出了 BART 模型在文本理解、摘要生成任务上的优越性。之后详细介绍了 BART-Large-CNN 模型，作为一个结合了 BERT 的双向编码器和 GPT 的自回归解码器优点的先进模型，BART-Large-CNN 特别适合于处理复杂的文本生成任务，如事件摘要和分析。该模型通过在大量新闻数据上的预训练，已经证明了其在生成精确、内容丰富的摘要方面的能力。

介绍完模型之后，详细讲解了本研究如何基于 BART-Large-CNN 实现金融舆情事件检测。这一部分首先简要介绍了本研究进行突发事件检测和事件摘要生成的流程，并且说明了本系统中如何通过使用 transformer 库中的 pipeline 函数完成与训练语言模型的调用并且指定任务类型为“summarization”。之后，通过与 PEGASUS-newsroom 和 T5 模型的对比实验，本研究进一步验证了 BART-Large-CNN 在处理需要详尽背景信息和深入分析的金融事件中的文本摘要生成任务中的优越表现。

本章的最后还展示了使用 BART-Large-CNN 进行金融舆情事件检测和摘要生成的实验成果，通过前端界面将分析结果以图表和词云的形式直观展示，使用户能够方便地获取和理解重要的金融舆情动态。这些功能的实现不仅提高了金融分析的效率，也为金融专业人士提供了强大的决策支持工具。

综上所述，本章围绕 BART-Large-CNN 模型的结构、原理以及在本研究中的使用流程及方法展开介绍，并且最后详细讲解了与 PEGASUS-newsroom 和 T5 的对比实验，得出了 BART-Large-CNN 在金融舆情事件检测任务上表现更出色的结论。

5、总结

本研究通过对事件检测技术的深入研究，成功基于预训练语言模型改进并完善了一个金融舆情事件检测系统，能够高效的完成数据收集、数据预处理、事件检测及摘要生成和分析结果展示等功能。在当今这个大数据时代，高效的数据分析、事件检测系统有助于相关从业者进行更高效、更全面的决策，特别是在金融领域，这样一个完善的系统能够极大的帮助投资者、经营者等了解市场动态、把握市场变化趋势，从而结合自身经验做出最合理的决策。

在数据采集阶段，本研究主要实现了从 Twitter 上获取指定的、与时事金融领域相关的关键词的大量推文，并且通过设计好的数据预处理流程，包括去除重复和无关信息、去停用词、去标点、格式化等步骤，确保了数据的有效性、完整性和准确性。在采集过程中，由于面对 Twitter 的反爬机制，针对性的设置了相应的处理机制，最大程度的减少了反爬机制对获取数据的影响。完成数据的采集和预处理后，还会对推文文本进行初步的分析，这一步包括完成关键词抽取、推文突发等级评估两个任务。基于知识抽取的关键词算法能够提取出每一条推文中的主谓宾三元组（如果没有则用 hashtag 和 user_mention 来补充）作为推文的关键词，之后又根据每条推文的三个关键词出现频率来判断该推文的突发程度。起初制定的评估规则是使用三个关键词的出现频率之和，但经过实践发现这会导致“特别重大事件”异常多，因此改用三个关键词出现频率的平均值，最终得到了较为合理的事件等级分布情况。

在事件检测、分析阶段，本研究通过在后端部署 BART-Large-CNN、PEGASUS-newsroom 和 T5 这三个预训练的语言模型，成功且较为高效的实现了突发事件检测、关键词权重计算和事件摘要生成等核心功能。最终，数据的分析结果在前端网页上通过折线图、词云、堆叠图、表格和时间线等多种方式进行展示，直观地给用户展现了数据的分析结果，使用户能从网页中直接获取有助于进行投资选择、交易决策的信息，可以帮助用户减少获取数据、分析数据的时间，进而提升决策效率和合理性。

展望未来，在本研究的基础上，可以进一步拓展数据的来源，例如中国新闻网、《泰晤士报》（Times）和 CNN 等权威媒体，来获取更全面、更多元的数据，从而提升数据的多样性和全面性。在前端网页，可以进一步丰富、细化数据的展示方式，例如引入实时数据更新和个性化的可视化仪表板等。此外，语言模型也在不断进行更新迭代，可以不断尝试更新、表现更好的语言模型作为系统的内核，进一步提升系统的效率。为了更好的给投资者提供建议，还可以考虑在事件摘要的基础上进一步输出投

资建议。这些改进可以使用户更全面、更精确的把握全球金融市场的动态，在海量的数据中高效的提取出有价值的信息，并及时的做出反应。

另外，由于时间和资源的限制，本研究未能对采集到的数据进行人工标注和编写人工摘要，因此缺少了参考摘要，导致无法使用标准的量化评估方法对模型生成的摘要的质量进行评估和比较。未来的研究可以考虑在数据集上进行人工摘要的编写，并使用 ROUGE[14]、BLEU[15]和 ME-TEOR[16]等指标来量化的比较每个模型生成的摘要的质量。这些指标能够从不同的角度评价摘要的准确性、完整性和流畅性，从而提供更全面的模型性能评估。

参考文献

[1] 华东师范大学. 基于关键词聚类的突发事件检测方法[P]. CN, A, ZL202011498455.2. 2021-03-19. https://pss-system.cponline.cnipa.gov.cn/retrieveList?prevPageTit=changgui.

[2] Han S, Hao X, Huang H. An event-extraction approach for business analysis from online Chinese news[J]. Electronic Commerce Research and Applications. 2018, 28: 244-260.

[3] Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced Language Representation with Informative Entities[C]. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019, 1441-1451.

[4] Liu Y, Yang D, Wang Y, Liu J, Liu J, Boukerche A, Sun P, Song L. Generalized Video Anomaly Event Detection: Systematic Taxonomy and Comparison of Deep Models[J]. ACM Comput. Surv., 2024, 56(7): 189, 38 pages. https://doi.org/10.1145/3645101

[5] Li Q, Chao Y, Li D, Lu Y, Zhang C. Event Detection from Social Media Stream: Methods, Datasets and Opportunities[C]. 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, 3509-3516. doi:10.1109/BigData55660.2022.10020411.

[6] Zhang J, Zhao Y, Saleh M, Liu PJ. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization[C]. International Conference on Machine Learning (ICML), Virtual, 2020.

[7] Raffel C, Shazeer NM, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[J]. J.Mach. Learn. Res., 2019, 21: 140:1-140:67.

[8] Stoyanov V, Zettlemoyer L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension[C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020, 7871-7880.

[9] Yenduri G, Ramalingam M, ChemmalarSelvi G, Supriya Y, Srivastava G, Maddikunta PK, DeeptiRaj G, Jhaveri RH, Prabadevi B, Wang W, Vasilakos AV, Gadekallu TR. Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions[E]. arXiv:2305.10435, 2023.

[10] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. In Proceedings of the 2019 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019, 4171-4186.

[11]Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit[J]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, 55-60.

[12]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C].NIPS, Long Beach, California, USA, 2017.

[13]Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, von Platen P, Ma C, Jernite Y, Plu J, Xu C, Le Scao T, Gugger S. Transformers: State-of-the-Art Natural Language Processing[J]. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, 38-45.

[14]Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries[C]. In Text Summarization Branches Out, Barcelona, Spain, 2004, 74-81.

[15]Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a Method for Automatic Evaluation of Machine Translation[C]. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, 311-318.

[16]Denkowski M, Lavie A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language[C]. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland, USA, 2014, 376-380.

致谢

在华东师范大学度过的四年时光，我由懵懂稚嫩逐渐变得成熟懂事，而在这个过程中，我得到了许多老师、同学和朋友的帮助和鼓励。光阴荏苒，日月如梭，转眼间我在华东师范大学的时光就走到了尾声。因此，在这篇毕业论文的最后，我要向这段旅程中给予过我帮助的人们、特别是在这篇毕业论文完成过程中给予我指导、支持的老师和学长们，表达我衷心的谢意。

首先，我要感谢我的毕业论文导师王晓玲老师。感谢您在整个研究过程中对我的悉心指导和支持。特别是在论文最后的写作环节，您传授给我的宝贵经验让我受益匪浅。这些经验不仅让我顺利的完成了这篇毕业论文，我相信也会在我之后的学习工作中不断发挥积极作用、帮助我不断进步。

其次，我还要衷心感谢在这次毕业设计的研究过程中，一直帮助指导我的杜威、张振宇和戴思龙三位学长。感谢你们在整个研究过程中的帮助、支持和鼓励。是你们在我陷入迷茫时为我指明了方向，在我研究陷入困境时帮助我解决了难题。因为有你们的支持和帮助，我才能顺利的完成这次毕业设计的研究。

此外，我要感谢大学四年里所有教导过我的老师们。是老师们的教导带我走进了计算机学科的大门，让我充分地积累了专业知识和技能。我要特别感谢兰曼老师和陈琴老师在我申请研究生的过程中给予了我充分的支持和帮助，让我能在自己感兴趣的领域进行更深入的学习和研究。

另外，我要感谢一路以来一直陪伴、支持我的父母和家人。你们是我最坚实的后盾，让我能毫无忧虑的专注于学业，让我有了不断进步、不断成长的土壤。我会在未来的日子里始终尽自己最大的努力提升自己，来回报你们的陪伴和支持。

最后，我要感谢一个特别的人。你在我陷入最低谷时出现在了我的生命里，用你的温暖和阳光帮助我走出了阴霾、重新找回了前进的动力。我们彼此陪伴、一同面对学习、生活中的困难和挑战，一起变得更加成熟、强大。我一直很遗憾没能早点与你相识，但我也很庆幸我们没有彼此错过。愿我如星君如月，夜夜流光相皎洁，希望未来的人生里，我们始终携手共进，奔赴山海。

在本科四年在华东师范大学的学习过程中，还有许多朋友给予了我无私的帮助和支持，在此我无法一一列举，但我仍然要向你们表示最诚挚的感谢，与你们的友谊让这段时光更加珍贵。也希望未来的岁月里，即便我们天各一方，友谊也能长存。

再次向所有给予我帮助的人表示最衷心的感谢！